

DOCUMENT RESUME

ED 271 124

IR 051 561

AUTHOR Whittaker, Robert
 TITLE Databases: Beyond the Basics.
 PUB DATE 28 Dec 85
 NOTE 14p.; Paper presented at the Annual Meeting of the American Association of Teachers of Slavic and East European Languages (Chicago, IL, December 1985).
 PUB TYPE Viewpoints (120) -- Speeches/Conference Papers (150)
 EDRS PRICE MF01/PC01 Plus Postage.
 DESCRIPTORS *Databases; Humanities; *Information Retrieval; Information Storage; Language Teachers; Microcomputers; Position Papers; *Research Tools; *Slavic Languages
 IDENTIFIERS *Bibliographic Data Bases; Nonbibliographic Databases

ABSTRACT

This presented paper offers an elementary description of database characteristics and then provides a survey of databases that may be useful to the teacher and researcher in Slavic and East European languages and literatures. The survey focuses on commercial databases that are available, usable, and needed. Individual databases discussed include: the Modern Library Association International Bibliography; the Library of Congress Catalog; the Arts and Humanities Citation Index; and the Research Library Information Network (RLIN). The paper concludes with comments on individual database concepts and microcomputer applications for research, writing, and teaching. (THC)

 * Reproductions supplied by EDRS are the best that can be made *
 * from the original document. *

ED271124

IR051561

DATABASES: BEYOND THE BASICS

Robert Whittaker

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)

☒ This document has been reproduced as
received from the person or organization
originating it.

☐ Minor changes have been made to improve
reproduction quality.

• Points of view or opinions stated in this docu-
ment do not necessarily represent official
OERI position or policy.

This will not be about basic principles of databases, or indeed about database theory at all. Rather I will comment on some database resources that are available for the teacher and researcher in Slavic and East European languages and literatures.

I assume you have a general idea of database structure: a database contains files, which in turn contain records, which in turn contain fields, which themselves actually have (or are) the data. The telephone book is a simple database: there are two files -- the white and yellow pages; each name in the book with its attached information is a record; the fields of the record are its parts -- last name, first name, address, number; records in the yellow pages have additional fields -- category of sales or service, advertising text. Some of the fields are of uniform size: the telephone number is seven characters; some are variable, like the last name. The two files share many records in common, but they are sorted differently--white by surname, yellow by category of service. The telephone book, or any card-index-type analogy is deceptive: its rigid layout deemphasizes relationships among the non-indexed fields of different records. A database is an unformatted mass of discreet information units, each with pointers to other units. The more powerful the management system, the more freely one can select among these units, the less constraint there is in sorting these items, and the more useful the database.

"PERMISSION TO REPRODUCE THIS
MATERIAL HAS BEEN GRANTED BY

R. Whittaker

What follows is a survey of databases available and useful specifically for Slavists with microcomputers and modems -- devices that allow the computer to communicate with a host system over the telephone. My observations mostly concern commercial databases: what is available, usable, and needed. I will conclude with some comments on individual database concepts for research, writing, and teaching.

The most common public and commercial databases are bibliographical. The best known are Dialog and BRS. These two firms make available on line electronic editions of a number of published bibliographies. The hand-out lists the more useful bibliographies from these services -- the MLA International Bibliography, The Library of Congress Catalog, The Arts and Humanities Citation Index, and Dissertation Abstracts. Some more specialized, but still useful bibliographies are also listed.

These two services are not inexpensive: right of access costs \$50 to \$75, for which one receives a password and an instruction manual--sometimes enormous, as in the case of Dialog. (If you need to buy a modem and telecommunications software [a program to manage your requests and the results of your searches], the cost to get started increases \$200 to \$400.) However, with the exception of the BRS/After Dark service, you only pay for what you use, and there are no minimums (except the minimum for CALICO, which is quite reasonable). After this initial investment, there are three types of costs: for the network line, for access time, and for records retrieved. All

the files and services have access charges, from \$.23 to \$1.50 per minute. When you request records to be shown on screen, there may be an additional cost per record. Finally, there are telephone charges, usually to local network numbers (Telenet, Dialnet, Uninet, etc.) which are cheaper than long distance.

You can do a good deal of searching for a few dollars. The trick is to plan and type in your search requests on your own (free) time, using a macro program. Then you log on, fire off your macros, capture the responses in memory, then get off quickly when you're done. In two or three minutes you can make a dozen searches and have some of the results typed, all for a total cost of under \$5.00.

The scope of the electronic database files is the same as that of their published versions. The principal limitations of electronic editions are date of origin (degree of retrospection), currency, and the time of the last update. These details are summarized also in the handout.

If your search for bibliographical information follows the format of the printed edition, it makes little sense to use the on line edition, e.g., if you want to find what articles were published on Belinskij in 1982, or if you want to know the title of V. B. Linsky's dissertation defended in 1984. The power of a database lies in searching for information without a fixed format: across a number of years, combining words/phrases in more than one field, or perhaps identifying publishers of translations of a particular type in a specified language. But there are some difficulties.

First, database file structures are ideosyncratic. Different files have different fields -- different types and names. Searching within title fields, or authors, language, abstract, descriptors, etc. can thus be confusing, especially since the type of contents in similar fields can differ among files. A multiple search for terms with no limitation to specific fields tends to collect unneeded hits. The "Moscow" in the first search collected the occasional place of publication. And the high number of hits on Arts and Humanities reflects the peculiarity of this file's CR (Cross Reference) field, of which more in a moment. True, the structure of each individual file is described in detail in the reference material one receives. But life would be much simpler if there were more structural and taxonomic continuity among bibliographical files.

A second difficulty is the lack of a common thesaurus for searching. Just to search for an author and a year, or for a word or phrase in a title is a relatively straightforward task. A database is most powerful when searched by subjects and topics. Knowing the terms used to describe records -- the searchable terms of a descriptor or subject field, then, is necessary for an effective search. One such thesaurus is the Library of Congress Subject Headings list. But this set of topics has its limits: try looking for books on a topic in the stacks of a library. The MLA began work on a thesaurus and indexing system in 1978, which they have called the Contextual Indexing and Faced Taxonomy system -- CIFT (the MLA should be good at acronyms). This system indexes a literature entry, for example, by language, performance

media, period, individuals, groups or movements, genres, and works; it also describes the literary topic in terms of features, literary techniques, themes-motifs-figures, influences, sources, processes; and it describes scholarly processes by types of scholarship, methodology, theories, devices, disciplines, scholars, special types of documents. [Examples from the handout.] To select records according to this system, it is helpful to know what terms are "searchable." However, the CIFT thesaurus is not available to users, but only to the MLA's bibliographers. A simplified version is promised in a year or so; but it may not include all the terms that are used for indexing and are therefore searchable. Once the MLA thesaurus becomes available, however, a few problems still remain: the descriptor field has not been evenly created -- earlier records are not up to later standards. [Examples from the handout.] As a rule, searching more recent entries, and not only for topics and subjects, is considerably more productive.

The Arts and Humanities Index file has a field called CR for cross-references. This field lists the references in an article, typically authors or works cited in footnotes. [Unfortunately, not all published materials: my attempts to find the contents of Clarence Brown's Viking Portable Russian Reader foiled.] Examples of this type of record are given in the handout. The CR field can skew a search badly: it can be very large and contain terms unrelated to the record. But it can make specialized searches quite effective. The Arts and Humanities Index has no subject or topic field, and therefore one is left with names and

titles; whatever subject or topic results from a search is often accidental, although perhaps also useful.

There are other, less known bibliographical services available on line. One of the most powerful (and expensive) is a database operated by the Research Libraries Group (RLG), which includes the country's major university, private, and government research libraries. This database is called the Research Library Information Network (RLIN), and access to the database for other than member institutions is provided by an exclusive vendor, the Cooperative Library Agency for Systems and Services (CLASS) -- see handout. The files in the RLIN database can be accessed by microcomputers with standard telecommunications software: searching costs about what other commercial databases charge. RLIN files are particularly useful as a union catalog, for the records indicate which libraries hold the works. The files themselves comprise the card catalogs of member and user libraries plus the machine readable catalog of the Library of Congress, beginning with 1963 acquisitions. (I am not aware of the extent to which RLIN includes acquisitions retrospectively, i.e. before the system was adopted.) Because the initial costs are quite high, one might rather visit the nearest RLG member library to use their terminals. If you can get access to these terminals yourself, you will soon have the pleasure of searching for records in Cyrillic (originally January, now put ahead to April 1986). When the whole system becomes Cyrillic-capable this spring, it will apparently be using a keyboard standard suggested by AATSEEL. (What the encoding conventions are I do not know,

nor what this Cyrillic code will do for individual microcomputer users without special terminals.)

The Library of Congress is another case of enhanced searching capabilities at the library site itself. Terminals at the Library provide access to its two information retrieval systems, MUMS and SCORPIO. MUMS (Multiple Use MARC System) combines searches of the several MARC files into one computerized catalog, and includes titles still being processed. SCORPIO (Subject Content Oriented Retriever for Processing Information Online) offers more sophisticated searching as well as a bibliographic citation file that includes the contents of periodicals in areas of current affairs. Both systems use the LCCC (Computerized Catalog), which begins with 1968 acquisitions (1979 for Cyrillic). Neither are available, it seems, except in the Library itself.

The Wilsonline is a unique on-line database, for it has no start-up costs. A manual costs \$30 -- not expensive. These databases are electronic versions of the Wilson Company's series of directories and indices. A couple are noted on the handout: their particular limitation for Slavists is that the bibliographies cover only English-language materials.

Specialized database files are provided by CALICO, the Computer Assisted Language Instruction Consortium at Brigham Young University. Details of this database and its special files are given on the handout.

What is missing is any database specifically for Slavists. Certainly we could share our own bibliographical research, at least. It is the nature of the electronic revolution to save time by eliminating repetition: perhaps a modest service for the collection and distribution of bibliographic database files can be organized. I would be grateful for suggestions and offers of help or data: over this past year some colleagues and I have been exploring the possibility of such a service.

Databases are not only bibliographic. There are textual databases as well, or simply texts which have been transferred to electronic media. There is a project at Rutgers University to collect information on these: the Rutgers Inventory of Machine-Readable Texts in the Humanities. Inquiries of this inventory indicate that there are no Slavic texts in their listing. In fact, it is my impression that there is no current project to collect machine-readable Slavic texts or information about them. The catalog of the Oxford Text Archive lists unnamed "samples of narrative" by Leskov as their only entry of any Slavic texts. However, machine-readable texts are surely being created (or converted): the computer-assisted concordances published by Prof. John Shaw, for example, suggest that behind the indexes lie complete texts. It would be a shame to spend time converting a substantial text, if this has already been done. Yet as of now there seems to be no systematic means to avoid such duplication.

Another form of database is the electronic network. The collection, organization, and distribution of data in a network or electronic bulletin board are database operations. There are two networks that seem to offer possibilities for Slavists. The first of these is the Foreign Language Educators Forum (FLEFO), which was noted in a recent issue of the AATSEEL Newsletter. This Forum is operated by CompuServe Information Systems, one of the largest commercial network-communication-database systems (sometimes called Videotex). Access to CompuServe costs considerably less than that for the other database firms. [see the handout]. One of the organizers and operators of FLEFO is Prof. Jerry Ervin of Ohio State. The forum has been in operation for just a few months and is still in its formative stages. It has a collection of files containing information and messages on a variety of topics -- (1) Conferences/Events (listings of upcoming conferences and events, including calls for papers), (2) Job Bank (positions wanted and positions sought), (3) The Registry (a directory of organizations and resources), (4) The Profession (a current events section for legislative updates, etc.), and a series of special interest files for (5) Students, (6) Educators, (7) Translators, (8) Literature, and (9) Linguistics. There is also a conference scheduled twice a week with a channel for Russian. It will be interesting to see how FLEFO develops. (Of the seven hundred users, I counted a dozen or so registered with an interest in Russian.) However, the obvious handicap for Russian and other non-Roman-alphabet languages is that the service has no conventions for the use of special character fonts.

There is another network being developed: SovSet'. I have not been able to find out much about this project: it is being developed by Thane Gustafson at Georgetown; it emphasizes Soviet affairs; and (apparently) it also uses CompuServe as a host. This network is rumored to be designed for the managing and sharing of information about American research as well as for collecting data on the Soviet Union.

These beginning efforts in electronic networks should in time produce several useful services: directories of programs, funding sources, research in progress, as well as files of bibliographic information and other data on Slavic and East European countries and cultures.

In conclusion, I would like to share some observations on software that is particularly useful for the individual creation of and work with databases on a microcomputer -- in particular for the management of text databases, ones that contain large amounts of text. In such database files, a record will have fields with several pages of text, or the record will have one field which is an entire document. The difference between such text-bases and the usual microcomputer database lies in the size and varying length of the fields. Usually, in standard databases (dBase II, PCFile, etc.) each field in a record has a fixed maximum length, and thus each record in a file is the same size. In a text-base, the field and record size is variable. I know of only one program (at a reasonable price) that can deal with entirely flexible fields -- Notebook II. It is exclusively a

text manager, and will do practically no calculation, for it has no numerical fields. However, it offers sophisticated searching and sorting features and can output the retrieved results virtually in any format. The utility of this type of text-base manager is suggested by its name: the records can be annotated -- for example, a bibliography can contain extensive comments -- almost without limit. A record is limited to 30 K (15 pages), and a maximum of 50 fields; the file size is limited only by disk capacity. The built-in text editor is like WordStar, but is unable to do block moves. On the other hand, one can create records with a wordprocessor and import them into the database. The importation of records retrieved from on-line databases is likewise easily accomplished with this program.

Another concept related to the textual database is even more flexible: this is the indexing of each significant word in an entire set of documents. This type of text-base file consists of records, each of which is itself a textual document having only one field. Programs like ZyIndex and FYI 3000 index any number of documents -- disks full -- and then manage searches for words, phrases, or sets of words. They retrieve the documents containing these words, or display the words in contexts of a set size. The utility of such programs for teaching and research lies in the ability to find names and concepts in multiple files of notes or of any kind of text document. The limitation of these programs is the form in which the results of the search are delivered. Generally, the "hits" are identified as being in a document which then must be retrieved as a separate process.

Since a record is only one field, the whole record must be retrieved. The combination of wordprocessor and text-base has been achieved by the program Nota Bene (which is also one of the more versatile Cyrillic word processors). The FYI program was adapted to this wordprocessor so that each paragraph in a document is treated as a separate field. Any number of documents can be processed into a database of textual files, each of which is a record of varying length whose fields are the paragraphs in each document. The results of a search can be reviewed, paragraph by paragraph, and selected paragraphs (fields) from the document can be combined into a new file. This file can then be reviewed separately, or inserted into the current file which is open in the wordprocessor. The clearest use for such a text-base program is the reviewing and even incorporating of notes from reading or other sources into one's current writing. This text-base is also similar to larger databases in that it creates its own thesaurus of searchable terms -- a listing of all the significant words in the documents that have been indexed -- and this list can be searched itself. This type of software is eliminating the distinctions between word- and text-processing and between data- and text-bases.

I would like to conclude with a comment on my own explorations along these lines. Databases or text-bases are conventionally alphanumeric. Microcomputers have another dimension beyond text and numbers -- namely graphics. Why not a graphics-database? I have been interested in the cultural topography of Moscow -- the image of Moscow in literature and art

and the relation of perceptions of the city to its actual economic, social, demographic, and geographical history. Locales, structures, events, works, and individuals all have temporal and spacial relationships. A combined alphanumeric and graphics database can, I suspect, demonstrate spacial relationships in two dimensions, and can overlay on this a third, temporal dimension. The eventual result would be a series of cultural maps, whose features change over time. One can search this database using alphanumeric terms (names, phrases, of dates) or using graphics data (locations, or areas). Perhaps, if perserverance and funding hold out, this research will produce insights into the cultural history of Moscow by means of a database that creates computer assisted guided excursions through the city in both space and time.

delivered 12/28/85

Annual AATSEEL Meeting

Chicago, ILL

--Robert Whittaker

Lehman College, CUNY